DOCUMENT RESUME

ED 125 504 HE 008 099

AUTHOR	Milton, Ohmer; Edgerly, John W.
TITLE	The Testing and Grading of Students.
INSTITUTION	Change Magazine, New Rochelle, N.Y.
SPONS AGENCY	Ford Foundation, New York, N.Y.
PUB DATE	76
NOTE	61p.
AVAILABLE FROM	Change Magazine, NBW Tower, New Rochelle, New York 10801 ($2.95)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	*Achievement Rating; *Achievement Tests; Bibliographies; Cognitive Objectives; Essay Tests; *Grades (Scholastic); *Grading; *Higher Education; Learning Motivation; Measurement Goals; Measurement Instruments; Test Construction; Test Reliability; Test Results; *Test Validity

ABSTRACT

Although over 100 million tests are administered each year and testing is a subject of increasing contention among students, faculty members remain diffident. A better understanding of the purpose and structure of evaluating mechanisms is a prerequisite for widespread improvement. Teachers must understand what factors play a part in measuring learning. If learning goals and course objectives are properly defined, they will be essential ingredients of success for student and teacher alike. Since multiple-choice and essay tests are most commonly used in college today, a thorough analysis of their structure and purpose is undertaken to clarify underlying principles of evaluation as a learning tool. Letter grading, the most commonly accepted form of evaluation, is particularly susceptible to the charge of insufficient feedback to the student. A more fundamental grasp of the options for academic measurement is the most direct route to improved grading. Growing external pressures are forcing faculty to reexamine student evaluation. The use of external examiners and the establishment of effective campus grievance arrangements are only two of the ways recommended to improve an increasingly bothersome issue in academic life. (LBH)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED125504

2.95

A Change Publication

The Testing and Grading of Students

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY
Change magazine
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

HE 008 099

By Ohmer Milton and John W. Edgerly

Library
National Center for Higher Education

The Change Policy Papers

1. Faculty Development in a Time of Retrenchment
2. Colleges and Money
3. **The Testing and Grading of Students**

Copyright © 1976 by *Change* Magazine and Educational Change
All rights reserved
Printed in the United States of America
LC 76-293
ISBN #0-915390-05-1

The Testing and Grading of Students *is one of a series of policy papers to help American faculty become more effective professionals. This volume has been published under a grant from The Ford Foundation.*

About This Special Report

WHEN VIEWED AGAINST THE CURRENT CRESCENDO OF egalitarian sentiments—or at least egalitarian rhetoric—the subject of student testing, let alone its design, leads one to some fascinating formulations about the very essence of an education. But no matter what one's ideological stance, the point remains that the "world outside" metes out rewards and penalties pretty much according to one's competence and talents. This being the current state of affairs, the least one can hope for is a less diffident effort by academics everywhere to evaluate student performance according to the fairest criteria available. There is now abundant evidence—and not only gleaned from student dissatisfaction—that testing and grading are often dispensed with an arbitrariness worthy of a Kublai Khan.

An appropriate design for tests should confirm the essential understanding between faculty and students as to what should be learned and what should not. This achievement is difficult, because it means, as Yale professor A. Barlett Giamatti has put it, "deciding that it is in fact a limited world, that some things are more important than others, that adjustments realistically have to be made. It means deciding that you really know what it is you want to teach and learn."

It is this centrality to the learning process that makes the subject of *Change*'s third policy paper, on testing and evaluation, so timely. The editors of *Change* are particularly grateful to the paper's authors, Ohmer Milton and John Edgerly, for their clearheaded portrayal of what is by any measure a vexing and complex subject. The authors are with the Learning Research Center and the Counseling Center at the University of Tennessee at Knoxville, respectively, and are widely regarded as sensitive authorities on the subject of human assessment. The preparation of their manuscript was facilitated by funds from the American Psychological Foundation, the Ford Foundation, and the University of Tennessee. This *Change* publication has been made possible under a separate Ford Foundation grant, which we acknowledge with thanks.

A number of individuals and organizations have contributed to the final formulation of this policy paper. The authors and editors wish to thank them for their counsel on a difficult and much debated

subject. They include John Bevan, College of Charleston; Kenneth Eble, University of Utah; John Gillis, Chapman College; Linda Kahan, Evergreen State College; Lee McDonald, Pomona College; Robert O'Neil, Indiana University at Bloomington; Robert Van Waes, American Association of University Professors (AAUP); Francis J. Wuest, Association of American Colleges (AAC); and Norman Frederiksen and Paul Diederich of the Educational Testing Service. Prior to publication, the AAUP and AAC endorsed this policy paper for its serviceability.

There is, on the editors' side, only one further wish: that this publication will be studied by thousands of faculty with as much care as was put into its preparation. One need not agree with every nuance and every thought expressed here: One need only be open to the possibility for learning much about the neglected subject of student evaluation. Here is as good a starting point as any to bring rational planning into what remain, surprisingly, still rather uncharted waters of academic life.

George W. Bonham
January 1976

eight

The Testing and Grading of Students: Why, Where, and How?

1

The Malignancy of Testing

Throughout American higher education, over 100 million tests are administered each year. Although testing is a subject of increasing contention among students, faculty members remain diffident. But a better understanding of both the purpose and structure of evaluating mechanisms becomes a prerequisite for widespread improvement. Page 11.

2

Setting Learning Goals

Teachers must understand what factors play a part in the measurement of learning. Faculty who pay ample attention to course content are often vague about the process of evaluation. If learning goals and course objectives are properly defined, they will be essential ingredients of success for student and teacher alike. Page 19.

3

Constructing Tests

Faculty widely confuse concepts of measurement and student evaluation. Regardless of the test construction chosen, both concepts must be carefully kept in mind. Multiple-choice and essay tests are most commonly used in college today. A thorough analysis of their structure and purpose clarifies underlying principles of evaluation as a learning tool. Page 27.

4

Grading

A comprehensive evaluation of student performance should provide guidance for academic improvement, but students too often receive scant critical commentary on their progress. Letter grading, the most commonly accepted form of evaluation, is particularly susceptible to the charge of insufficient feedback to the student. A more fundamental grasp of the options for academic measurement is the most direct route to improved grading. Page 43.

5

Lone Efforts Are Not Enough

Growing external pressures are forcing faculty to take a fresh look at student evaluation. The new consumerism, recent legal decisions, and far-reaching social criticism will no longer leave matters of grading and testing to the private academic preserve. The use of external examiners and the establishment of effective campus grievance arrangements are only two of the ways recommended to improve an increasingly nettlesome issue in academic life. Page 49.

6

For Further Reading

For a more comprehensive understanding of testing and evaluation, faculty have access to a number of excellent source documents. Here are some of the best. Page 57.

1

The Malignancy of Testing

Throughout American higher education, over 100 million tests are administered each year. Although testing is a subject of increasing contention among students, faculty members remain diffident. But a better understanding of both the purpose and structure of evaluating mechanisms becomes a prerequisite for widespread improvement.

THERE ARE SLIGHTLY MORE THAN HALF A MILLION faculty members in American colleges and universities. If each teaches an average of two courses and prepares three tests for each course, at least three million exams are given during any quarter or semester. Since these examinations are administered to about 10 million students, about 30 million tests are given every three or four months, or over 100 million every academic year. This is measurement on a grand scale indeed!

Considering that major decisions are made about students' lives—whether they remain in school, enter professional or graduate institutions, secure jobs—partially on the basis of those haloed test statistics, the grade point averages, elaborate care should be required in the entire testing and grading enterprise. Unfortunately, the very terms "testing" and "grading" have come to be used more or less synonymously, with either one referring to the entire process. In this policy paper each term will be used in a restricted and distinctive sense. Here, "testing" means measurement; "grading" means assigning an evaluative symbol—A, B, C, D, F (see Chapter 3).

While there are no documented reports about the degree to which care is exercised, a number of factors indicate that too much academic measurement in the classroom is conducted in a cavalier fashion. On the basis of an inspection of numerous tests over the years, loudly voiced and sometimes embittered laments by many students, and observation of too many untutored graduate teaching assistants assigned the entire chore of testing and grading, we developed a healthy skepticism about the practices in force. To check these initial suspicions, two dozen college and university officials

were surveyed. In many instances, a disproportionate share of the complaints they were receiving concerned testing and grading. (The range was from 2 to 80 percent, with the overall average percentage near 30.) Discussions with students and our faculty colleagues revealed additional evaluation problems.

The following cases exemplify many of the responses we received.

Complaints About Test Content

(1) **In an introductory course covering approximately 2,500 years of philosophy, five of the seven questions on the final exam were about Kant.**

Good testing should be based on representative sampling. Five of seven questions focusing on one philosopher within a period of 2,500 years is not adequate sampling.

Considering that major decisions are made about students' lives—whether they remain in school, enter professional or graduate institutions, secure jobs—partially on the basis of those haloed test statistics, the grade point averages, elaborate care should be required in the entire testing and grading enterprise. Unfortunately, the very terms "testing" and "grading" have come to be used more or less synonymously, with either one referring to the entire process.

---

(2) **Five textbooks were assigned in a course designed to promote understanding of the contributions of several individuals to the field being studied. Lectures and class discussions focused almost entirely on one of these people. All questions on the final concerned the one person.**

Instructors need to be much more careful than many seem to be in correlating objectives, assignments, and testing.

(3) **Students in a senior course were assigned 15 journal articles, the shortest of which was 14 pages long. The only examination question over this considerable volume of material asked students to match the articles' authors with the titles. Challenged by a colleague, the instructor argued he could assume that a student who could do this matching understood the material.**

The assumption made by this professor requires a major leap in logic. Moreover, his future students might play the odds and limit their learning to memorizing authors and titles.

Complaints About Grading

(4) Students in a technology course were assigned a project to be completed individually. They were informed that the only grading criterion would be the quality of the product. Several students devoted many hours to the assignment early in the term and finished several weeks before final exams. Other students maintained a leisurely pace and worked throughout the term. All the early finishers received F's, while the leisurely ones received A's and B's. The instructor maintained that many of the F's were awarded because of absences—he had not seen some of these students during the last month of the course. Even when reminded that his announced criterion was quality, not class attendance, he refused to alter the F's.

Many teachers are unduly touchy about class attendance. To be fair and just, measurement should measure what has been asked for and nothing else.

(5) A foreign language class contained students who had lived abroad, graduate students working toward the foreign language exams, and beginners. One error of any kind reduced a quiz or exam grade to a B, two errors to a C, and so on. Many studious and responsible students received D's and F's in this course.

While high standards are important, there should be some demonstrable relationship between reasonable expectations and grading criteria.

(6) After assigning quite high mid-term grades, an instructor declared that students were being coddled. Without any warning, term papers were graded very harshly (after the withdrawal date had passed). The final was graded equally severely. Course grades for the class of over 40 included one A, one C, and a few D's; the rest were F's. Colleagues of the instructor arranged for the students to take another final.

Capriciousness and arbitrariness have no place in evaluation.

(7) A freshman was told by an instructor she would receive a B in his course; her final grade was C. The student was applying for admission to a competitive program for which a few hundredths of a point in her GPA might determine her acceptance. Investigation revealed that the instructor was a teaching assistant who had left school. The department chairman believed this instructor's teaching and testing methods had been questionable and changed the grade. The student was admitted into the competitive program.

Most teaching assistants do not receive formal instruction or guidance about testing and grading. Incidents such as this one may be prevalent.

(8) In a course which had a fairly rigid attendance requirement, a student requested an excused absence for a

mandatory appearance in court and believed the instructor allowed the absence. The student received a final grade of C instead of the B he had expected. The instructor explained the C in different ways: to the student—too many absences (he disallowed the court appearance); to the department chairman—inadequate class participation (records indicated otherwise); to an administrator—poor written work (papers averaged B). The instructor refused to change the grade, but an academic grievance committee directed a change.

If an instructor justifies a grade in so many ways, how can evaluators of the student's transcript interpret the grade?

(9) A freshman who had maintained a C average on all his tests received a final grade of F. The instructor explained that the student had exhibited an improper spirit toward the subject matter and refused to alter the grade.

We have serious doubts about the propriety of grading a student's "spirit" or attitude.

While there are no documented reports about the degree to which care is exercised, a number of factors indicate that too much academic measurement in the classroom is conducted in a cavalier fashion. On the basis of an inspection of numerous tests over the years, loudly voiced and sometimes embittered laments by many students, and observation of too many untutored graduate teaching assistants assigned the entire chore of testing and grading, we developed a healthy skepticism about the practices in force.

---

(10) A student with the highest overall point total in her class (90 percent) received a B rather than an A as her final grade. The instructor explained that his point system was absolute and that while she had a "moral A," he could not give her an A for the course. During the conversation he told her he had given another student a B when that student had only enough points for a C. He rationalized that there was a difference between giving someone a B and giving someone an A but did not explain what the difference was. The student appealed to the grievance committee.

Undergraduates state that they frequently encounter this professorial attitude, although seldom in such a blatant incident. Assigning a grade in such a manner is not responsible evaluation.

(11) A female student was informed by a professor: "Women do not belong in my field." Her grade for the course was significantly lower than the average she had maintained.

The injustice was rectified with the assistance of the department chairman.

Social prejudices must be eliminated in evaluating student achievement, and every effort should be made to minimize personal prejudices.

Complaint About Test Conditions

(12) **A final was administered to 130 students in a crowded classroom where it was easy for students to copy from each others' papers. Although many students thought the situation was unjust, the instructor refused to change the test location, blaming the institution for assigning too many students to the class and for providing the small classroom. An administrator, the department chairman, and the college dean intervened, and the test was given again under satisfactory conditions.**

Inadequate and improper testing conditions are inexcusable.

Most student complaints seem to be about grades or the symbols, not about testing or measurement where the basic problems are. Apparently most students are unaware of the fundamental issues in measurement and evaluation and do not know the questions they should be asking. They are not alone in seeing just the tip of the evaluation iceberg. Thousands of studies have been conducted about grades and grade point averages [GPAs], but the measuring devices from which those symbols are derived are rarely questioned.

Most of the students' complaints seem to be about grades or the symbols, not about testing or measurement where the basic problems are. Apparently most students are unaware of the fundamental issues in measurement and evaluation and do not know the questions they should be asking. They are not alone in seeing just the tip of the evaluation iceberg. Thousands of studies have been conducted about grades and grade point averages (GPAs), but the measuring devices from which those symbols are derived are rarely questioned.

Study Influences

The effects of testing upon learning have been almost totally ignored, yet experimental scientists have been concerned for many years with the effects the act of measurement has upon the object or phe-

nomenon being measured. A good example is a blood-pressure reading. At least two features of the act of measuring blood pressure distort the true reading: the pressure of the inflated cuff and, for some people, the emotional reaction to the procedure. The reading is false to some degree because of either or both of these.

Generations of students have told their faculty that testing influences them. They study according to the type of test they are going to take and in so doing learn different features of the material. A few studies support their assertions. Meyer found, by analyzing notes made and the booklets which contained new material to be learned, that a smaller percentage of students who were to receive an essay test used underlining and a greater percentage of them made summaries than students who were to take objective tests. Thomas and Augstein found that students who were informed that their test on a paper on genetics would be in essay form, but who in fact took objective and essay tests, performed better on both types than did students who studied the same material under the impression that their test would be objective (but received the two types). Felker and Dapra demonstrated that comprehension-type questions were more effective for enhancing problem solving than verbatim-type questions.

Directions

It seems likely that traditional testing and evaluation practices—written tests covering subject matter and grading on curves—will continue on a grand scale, especially in lower-division courses. The remainder of this policy paper is devoted primarily to introducing faculty members to basic principles of measurement and some of the prominent unresolved issues of grading. Improved testing devices and practices will help learning and grading. Numerous volumes have been written about most of the topics that we only mention; carefully selected references are given in Chapter 6. The purpose throughout this volume is to alert faculties to some exceedingly complex problems: The measurement of learning, the assigning of grades, and determining the significance of the process are inordinately complicated procedures.

2

Setting Learning Goals

Teachers must understand what factors play a part in the measurement of learning. Faculty who pay ample attention to course content are often vague about the process of evaluation. If learning goals and course objectives are properly defined, they will be essential ingredients of success for student and teacher alike.

ONE OF THE STUDENT-REACTION-TO-INSTRUCTION FORMS used at the University of Tennessee in Knoxville allows instructors to write in extra items. Just before a term ended, one instructor wrote in, "Rate your progress on the course objectives." We suggested that he might list the objectives himself and ask the students to rate their progress on each one. He replied he wasn't certain what the objectives were, but he would try to determine them after the course was over.

How does one measure at all if one does not know what one wishes to measure? Instructors should not be like St. Augustine when he declared, "For so it is, O Lord my God, I measure it; but what it is I measure, I do not know."

Goals and Objectives

It should go without saying that effective evaluation (testing and grading) is based on well-established goals and objectives, yet frequently it is not. Faculty devote great amounts of attention to the content of their courses (what to include, how to include it, and what to exclude), but too few give as much time or energy to the process of evaluation, even though the goals and objectives of a course and of evaluation are the same.

Goals and objectives are often thought of separately so that their roles in evaluation may be delineated. Goals may be defined as the hoped-for end results or products of a sequence of educational events. Goals may apply to a single course or to a sequential pro-

gram (e.g., a major). Objectives are the short-range events in a sequence leading to a goal.

Goals can best be measured through the assessment of well-defined objectives. This principle is as true for a professor's course as it is for a college's curriculum. The goal for this volume is "improving testing and grading." One measure might be the number of faculty who seek to apply the principles expounded. The objectives are for faculty to understand the detailed ways of attaining the goal. One measure might be their performances on a carefully constructed written test over the contents.

Goals and objectives should be stated in as empirical a fashion as possible so that they will be susceptible to evaluation. It is true that some educational goals are difficult to state in definitive terms, but difficulty is no excuse for not trying to come to grips with the clarity of goal statements.

True, some curriculum and course goals do seem to defy evaluation. Such goals, often found in college catalogs and course syllabi, usually run as follows: The liberal arts education provides the individual with the ability to comprehend the great outlines of knowledge, the principles upon which it rests, the scale of its parts, its

---

It should go without saying that effective evaluation [testing and grading] is based on well-established goals and objectives. Yet frequently it is not. Faculty devote great amounts of attention to the content of their courses [what to include, how to include it, and what to exclude], but too few give as much time or energy to the process of evaluation, even though the goals and objectives of a course and of evaluation are the same.

---

lights and shadows. A liberally educated person is identified by quality of mind. Educators insist these respectable and cherished goals should not be compromised. As stated, they correspond to the accepted definition of a goal as an abstract statement of a hoped-for result (Mager, 1972). They do not, however, tell how to achieve results. This is where objectives play a crucial role in describing what knowledge, skills, understanding, and behaviors (such as laboratory abilities) the students should possess after completing their experience of the curriculum.

It is in defining objectives that many courses and curricula fall short and thereby complicate evaluation. It is generally assumed that the lauded goals are accomplished through various curricula, but the objectives are stated no more clearly than the goals, and hence the confusion.

Although this presentation of the basic principles of setting goals and objectives is concerned with the level of the individual course and the individual test, what pertains at this level is applicable to an entire curriculum. Courses within a curriculum are assumed to be cumulative. The vast majority of courses have prerequisites that

assume that the successful completion of one course's objectives provides passage to the next course. It is assumed that all the courses contribute to the goals of a curriculum or program.

Matching Test Items and Objectives

One of the most frequent violations of good procedure in setting objectives for achievement assessment is a mismatch between the objective and the unit of measurement chosen to assess it. The basic unit of measurement of an objective is the individual test item and it is imperative that the two be well matched. In courses in measurement, though, even some bright and well-informed graduate students have great difficulty preparing test items that adequately match the stated objectives. Matching is difficult, but not impossible, particularly if objectives have been carefully stated.

As Mager (1973) so aptly states: "The issue of inappropriate test items is a widespread phenomenon. . .and a practice (mal-practice?) most urgently in need of improvement. When we deceive the student by discrepancies between our words and our deeds, both he and we are the losers."

The first task in testing, therefore, is to define objectives clearly. These should be made as concrete as possible. Then matching the unit of measurement (the test item) to the objective becomes somewhat easier and one can choose the appropriate test item format. If, for example, "knowledge of," as opposed to "skill in," an academic area is a course goal, then one would choose a compatible set of objectives and units of measurement (test items) to assess its achievement. These two quite different tasks obviously call for different performances or behaviors.

Students frequently complain that an exam did not cover the content of the course. This often means that there was a mismatch between the test items and the objectives. It is crucial to have a sound understanding of what types of performances are required by the objectives. Only in this way can one construct the appropriate item to measure the achievement of the objective. In this sense, each test item is a criterion-referenced item; that is, each item serves as a means of identifying a student's status with respect to an established standard or of assessing objectives.

We pointed out earlier that some academic fields appear to be more easily accessible to measurement than others. However, one encounters as many errors in tests assessing achievement in mathematics as in literature. In this regard, no academic domain or area seems to be entirely free of error. What appears to be a discrepancy between complex (difficult to get at) and simple (easy to get at) domains could be greatly reduced if the basic principles of setting objectives were followed. In short, one must make the decision whether the objectives will lead to "knowledge of," skills, "concepts about," "understandings of" (all manifested in writing) or overt behaviors (manifested by manipulating laboratory equipment, for example).

There is probably no better way of stating an objective (or initiating thinking about objectives) than to pose the following question at the outset of the course: What do I want my students to be able to do

at the end of this course?

This usually generates a long list of rather lofty goals that must be translated into objectives. The key words are *be able to do*: read a map, prepare a brief, test for diabetes, explain how a bill becomes a law, describe the human eye, solve an equation, write an essay.

Domain Dictates Objectives

There is perhaps nothing more frustrating to students than to be told that a course objective is for them to be able to write a grammatically correct theme of 100 words and then have a well-meaning professor discount points for lack of imagination and/or creativity. Errors of this sort are commonplace. Not only is this an error in stating objectives to students, it is also an error in the choice of the appropriate item format or type. Creativity is an extremely difficult

---

One of the most frequent violations of good procedure in setting objectives for achievement assessment is a mismatch between the objective and the unit of measurement chosen to assess it. The basic unit of measurement of an objective is the individual test item, and it is imperative that the two be well matched. In courses in measurement, though, even some bright and well-informed graduate students have great difficulty preparing test items that adequately match the stated objectives. Matching is difficult, but not impossible, particularly if objectives have been carefully stated.

---

area to assess—but not impossible. One first defines it and then chooses or constructs the appropriate items to assess its presence or absence.

The problem of choosing appropriate objectives and subsequent test items for assessing achievement within a given domain or area does point up that the domain determines objectives, to a degree. We are not suggesting that faculty back away from trying to assess those goals that they regard as important just because a goal might seem fuzzy. We are not sympathetic with those who contend that the crucial things within their domain are inaccessible to objective assessment and who often claim that only experience and subjective judgment can serve as bona fide assessment. We repeat: If something is worth being made a goal, it is worth being objectified! This position in no way lessens the admirable qualities of a goal.

A good example is a course in art appreciation. If a goal of the course is to appreciate fine art, one simply has to state what the student should be able to do at the course's conclusion. For

example, the student might be expected to be able to choose from a list of paintings five that would be considered as representative of fine art by a panel of experts, fine art, in turn, having been made definable by excluding from it violations of characteristics common to fine art, e.g., good composition, perspective, and so on.

The claim that these types of assessment issues are not accessible or are too open to subjective interpretation is inaccurate. There is little question that today's dime-a-dozen novels will not be tomorrow's literary masterpieces or Pulitzer Prize winners. There is little room for doubt that Dante's *Inferno* is superior. Subjectivity enters when one is asked to indicate whether one likes or dislikes a book. This is a personal rendition of one's own experience, but to be able to discern the characteristics of great literature from a random selection of books is something someone can learn to do and subsequently demonstrate.

To emphasize the importance of defining and specifying performance objectives, Mager (1973) suggests the rather humorous "Hey Dad" technique. Here, one places a course objective within the following context: "Hey Dad, let me show you how I can ____!" If the result of filling in the blank is a seemingly absurd statement, the objective is too broad and needs clarification and simplification. In our example of art appreciation, as a course objective, the following absurdity would be the result: "Hey Dad, let me show you how I can appreciate fine art!" This absurdity can be obviated by specifying the generally agreed-upon component behaviors or performances of art appreciation. The following examples make the initial objective more tolerable:

> Hey Dad, let me show you how I can, when presented with them, accurately identify 10 out of 10 Renaissance paintings, supply their titles and the artists' names, name two additional paintings each has done, when and where each lived, three contributions each has made to the history of art, and two elements of their work that have led them to be judged as outstanding in the history of art.

In this fashion, art appreciation becomes less fuzzy and is more easily assessed.

An Illustration

There are several ways of measuring the extent to which course objectives have been met. As previously mentioned, the domain or area does exercise some influence over the type of test or measurement one uses to assess course objectives. There is, however, a basic reciprocity between the types of test employed and the objectives of a course. For example, it just makes good sense to use performance (i.e., observable behavior) to assess the objectives of performance courses. Most of the physical sciences require laboratory skills, the attainment of which requires the instructor to observe whether the student can do the task in question. Most academic courses are assessed by asking students to perform on a written exam. In other words, instructors are assessing students' ability to do something vis-à-vis their response to a written question. Within this form of testing, we ask them to demonstrate knowledge of or about

in a variety of ways: multiple-choice testing, matching, true-false, and essay, to mention just a few of the varieties.

An example demonstrates how a simple objective is amenable to different testing forms. The basic case of "Making a Pot of Coffee" is drawn from Mager (1973). Making a pot of coffee with an electric coffee pot calls for knowing how to do definite things:

> (1) Disconnect coffee pot; (2) disassemble coffee pot; (3) clean components and pot; (4) inspect components of pot; (5) fill pot with water; (6) reassemble components of pot; (7) fill basket with coffee; (8) reconnect coffee pot; (9) set dial on coffee pot; (10) note if pot is perking properly.

A student's knowledge can be assessed in a variety of test types. One of the objectives in teaching coffee making might call for a knowledge of (or ability to recognize or state) the correct sequence of action in making a pot of coffee. One multiple-choice question could take the following form:

1. Of the items below, which is the first step in making a pot of coffee:
   (a) fill the basket with coffee
   (b) note if pot is perking properly
   (c) disassemble coffee pot
   (d) *disconnect coffee pot*

An essay question requiring this same knowledge might take the following form: Please describe in no more than 100 words the 10 important steps in making a pot of coffee.

A matching test on coffee making might be prepared as follows:

Take every other step and make a comparison right and left list:

	Left		Right
Step	1 disconnect coffee pot	Step	2 disassemble coffee pot
	3 clean components and pot		4 inspect components
	5 fill pot with water		6 reassemble components
	7 fill basket with coffee		8 reconnect coffee pot
	9 set dial on coffee pot		10 watch to see if pot is perking properly

Then shuffle the right list to derive the following:

Step	1 disconnect coffee pot	Step___	reassemble components
	3 clean components and pot	___	disassemble coffee pot
	5 fill pot with water	___	watch to see if pot is perking properly
	7 fill basket with coffee	___	inspect components
	9 set dial on coffee pot	___	reconnect coffee pot

The matching test for the students would then be:

> The list on the left contains the correct ordering of steps 1,3,5,7, and 9 of the 10 appropriate steps in making a cup of coffee. The list on the right contains steps 2,4,6,8, and 10. However, the steps on the right have been shuffled. Your task is to draw a line from Step 1 on the left to the appropriate Step 2 on the right; a line from Step 3 on the left to the correct Step 4 on the right and so on until you have correctly matched all 10 steps in their correct sequence.

As we shall presently see, test construction is a time-consuming task, principally because the preparation of learning objectives must be done with great care. This is the key to successful testing.

3

Constructing Tests

Faculty widely confuse concepts of measurement and student evaluation. Regardless of the type of test chosen, both concepts must be carefully kept in mind. Multiple-choice and essay tests are most commonly used in college today. A thorough analysis of their structure and purpose clarifies underlying principles of evaluation as a learning tool.

OUR INVESTIGATIONS INTO STUDENT ASSESSMENT HAVE led to several conclusions: (1) There is real confusion about the concepts of measurement and evaluation. (2) Many faculty members believe their discipline is so unique that little is to be learned about academic measurement from faculty of other disciplines. (3) Instructors feel there must be no interference in their testing and grading of students—not even by their own disciplinary colleagues.

Tests should promote learning. They should assist the student and the instructor in determining whether learning goals are being achieved. If they do not, then both participants may alter strategies. In this private context, formal measurement is of little importance, because errors in judgment by the instructor can be corrected and honest differences of opinion can be resolved. Central to exchanges between the two is the student receiving detailed criticism of his or her work and constructive suggestions for improving it.

What has happened, however, is that the letter symbols resulting from tests are used almost solely for official record keeping. Many instructors do not view testing as part of the learning process and as a result resent spending class time on it, return exams to students with no correction marks or comments upon them, and never show final exam results to students. Students, in accepting this limited use of tests, strive to gain points rather than to learn.

In this context, it is difficult to understand how the defensive cry of "academic freedom" (meaning "Stay away; I'll test and grade as I please") can be justified. Faculty members are fallible. They can be capricious (Case 6, page 14) in their judgments of student achievement, and poorly constructed tests can support those judgments. In the final analysis, it is the student who pays the price; and

the best students are harmed the most. They are the ones who engage in "grade grubbing" because they hope to enter graduate and professional schools, and very tiny fractions of GPA points may decide their fates.

The thesis here is simple: Since the results of measurement of student achievement are currently used more to serve the public than to promote learning (that is, the results are made available to employers and others to be used in the selection process), individual faculty members can no longer pretend infallible judgment about student assessment. While we disagree with this public function, since it will continue it must be improved. This chapter will explain and clarify the concepts of both testing and grading and introduce some of the necessary principles for techniques of measurement.

Concepts: Measure, Evaluate

The word "measure" has at least 40 different meanings (Lorge). In the present context measure is intended to mean all those activities

---

Since the results of measurement of student achievement are currently used more to serve the public than to promote learning [that is, the results are made available to employers and others to be used in the selection process], individual faculty members can no longer pretend infallible judgment about student assessment. While we disagree with this public function, since it will continue it must be improved.

---

which are necessary to quantify learning or achievement: the preparation of single questions or items, the selection of items or questions to make up a test or examination, the conditions under which the test is administered, scoring each individual item, and assigning a score, number, or quantity to the whole. In everyday parlance, all of these activities are referred to as testing.

The goal of objectivity is sought in all measurement. In the hands of several trained people, the same instrument—whether a ruler, a watch, a sextant, a sphygmomanometer, an English test—should yield the same reading. Ebel's (1972) definition applies with equal force to all educational tests: "A measurement is objective if it can be verified by another independent measurement. If it cannot be, that is, if the measurement reported depends more on the person making the measurement than on the person being measured, it is unlikely to be very dependable or very useful...."

The greater the care with which an instrument is constructed, the greater the likelihood that two or more trained people will obtain the same reading (or quantity or score) for the same value or operation. Most people seem to be alert to this principle for physical

measurements, but much less attuned to it for educational ones. This general lack of sophistication is illustrated by the prevalence of superficial thinking about so-called objective tests. Multiple-choice and true-false tests are both called objective because two or more scorers will arrive at the same score for an examinee after a key is prepared. But the score or quantity assigned is only one aspect of measurement; if other principles of measurement have been applied carelessly, the test is not objective.

It is a common error to equate quantification, no matter how determined, with objectivity. As Hofstadter has explained: "The American mind seems extremely vulnerable to the belief that any alleged notion which can be expressed in figures is in fact as final and exact as the figures in which it is expressed." Upon reflection, it is clear that, for example, an 85 on a test paper could have been derived arbitrarily, and the instrument on the basis of which it was calculated could have been constructed poorly in the first place.

As we use the term, "evaluation" means arriving at a judgment or decision. The physician, after taking a blood pressure reading, makes a judgment that the blood pressure is normal or abnormal.

The greater the care with which an instrument is constructed, the greater the likelihood that two or more trained people will obtain the same reading [or quantity or score] for the same value or operation. Most people seem to be alert to this principle for physical measurements, but much less attuned to it for educational ones. This general lack of sophistication is illustrated by the prevalence of superficial thinking about so-called objective tests.

The driver, after trying to collect sound information about two automobiles, weighs the evidence and buys car A rather than car B. The instructor examines a student's test performance, reaches a decision about the level of achievement, and expresses it in a letter symbol. Needless to say, such decisions may not be simple in reality. While the goal in measurement is objectivity, one of the chief goals in evaluation is minimizing extraneous factors or variables. In Case 11 (page 15), the sex of the student was an extraneous factor and should have had nothing to do with her final grade (evaluation). Ultimately, evaluation is subjective because human judgment is its essence. The greater the extent to which judgments are based upon carefully constructed and administered measuring devices, the greater the likelihood they will be sound. Factors to be considered in evaluating student achievement (assigning grades) are discussed more fully in Chapter 4.

Test-Question Principles

There is at least one unalterable fact about testing: It is time-consuming. There are no short cuts to constructing a good test. Tests take many forms—multiple-choice, true-false, essay, matching, completion, problems, interpretive, and combinations of these. We have set forth certain principles and recommendations that are applicable to written tests because without question such tests are used almost exclusively in higher education. In this connection the work of Ebel (1966, 1972) has been drawn on heavily, and the reader might also see Adkins and Dressel.

The basic unit in a written test is the individual item or question—improvement in measurement begins at this point. Judging from the literature, less attention has been devoted to item preparation than to any other feature of test construction. For this reason, certain principles of item preparation are emphasized, with many examples.

Instructors who prepare items or questions must possess several abilities:

- *A thorough mastery of the subject matter.* Item writers must be acquainted with facts and principles, attuned to their implications, and aware of popular fallacies and misconceptions. Most graduate teaching assistants do not have such mastery.
- *A rational and well-developed set of aims or objectives for the instruction.* For most courses these will include helping students learn facts and principles, make abstract generalizations, be critical, and apply what has been learned in other settings. The importance of aims and objectives cannot be overstressed.
- *A mastery of written communication.* Those who have written for publication have learned how difficult it is to choose the right words and to arrange them to convey the meaning intended. Students probably give the words in test questions much more critical attention than almost any other prose receives.
- *A knowledge of the special techniques of item writing and how to use them.* Some of these will be discussed further on.

Since the two test forms used most commonly are multiple-choice and essay and since our space is limited, we will discuss the development of only these two in some detail.

Multiple-Choice Questions

Multiple-choice tests have been condemned roundly by many instructors and students (the latter sometimes refer to them as multiple-guess). Much of this criticism is well-founded because many tests are constructed carelessly. Items tend to be ambiguous and to emphasize the trivial. In one study (McGuire), three judges classified test items that covered knowledge in medical subjects and unanimously agreed that over half of the items measured predominantly recall and recognition of isolated information. Fewer than one fourth of the items were thought by any single judge to require even simple elements of interpretation or problem solving.

Properly developed, however, multiple-choice tests can tap many facets of learning. The principles here set forth are merely introduc-

tory and may appear deceptively simple, but their application is time-consuming and demanding. Illustrative questions or items* are uncomplicated in the hope they will enable the disciplinary specialist to focus upon the principle.

(1) *Strive for item clarity.* The English language is full of ambiguous words. The printed page cannot convey such clues to meaning as voice inflections and facial expressions. Test items should not be verbal puzzles. A test's purpose is to test or measure knowledge rather than verbal puzzle-solving ability. The major recommendation for attaining clarity in items or questions is: *Every item, before it is used, should be responded to by a colleague and by an advanced student (the latter will detect vagueness, ambiguities, and errors the former might miss).*

(2) *Include in the stem or body all necessary qualifications that are needed for answer selection.* Consider the following multiple-choice question:

> If a ship is wrecked in very deep water, how far will it sink?
> 1. Just under the surface.
> 2. To the bottom.
> 3. Until the pressure is equal to its weight.
> 4. To a depth which depends in part upon the amount of air it contains.

The instructor intended 2 as the correct answer, but several capable students chose 4 because they considered the possibility (which the instructor failed to exclude) that a wrecked ship might not sink completely.

(3) *Generally, omit nonfunctional words. They tend to interfere with comprehension.* Consider:

> While many in the U.S. feared the inflationary effects of a general tax reduction, there was widespread support for a federal community-property tax law under which:
> 1. husbands and wives could split their combined income and file separate returns.
> 2. homesteads would be exempt from local real estate taxes.
> 3. state income taxes might be deducted from federal returns.
> 4. farmland taxes would be lower.

Comprehension of this item may be facilitated by rewording it as follows:

> Community-property tax laws permit:
> 1. husbands and wives to split their combined income and file separate returns.
> 2. homesteads to be exempt from local real estate taxes.
> 3. state income taxes to be deducted on federal returns.
> 4. farmland taxes to be lowered.

Sometimes, though, it is useful to include introductory statements that help to emphasize importance:

> The pollution of streams in the more populous regions of the United States is causing considerable concern. What is the effect, if any, of sewage on the fish life of a stream?

*These are from Ebel, Robert L. "Writing the Test Item." In **Educational Measurement**, edited by E. F. Lindquist. Washington, D.C.: American Council on Education, 1966, and are used by permission.

1. It destroys fish by robbing them of oxygen.
2. It poisons fish by the germs it carries.
3. It fosters development of nonedible game fish that destroy edible fish.
4. Sewage itself has no harmful effect on fish life.

(4) *Beware of unessential specificity and/or trivia.* Consider:

What percent of the milk supply in municipalities of over 1,000 was safeguarded by tuberculin testing, abortion testing, and pasteurization?
1. 11.1 percent
2. 20.3 percent
3. 31.5 percent
4. 51.9 percent
5. 83.5 percent

This item, encouraging rote memorizing, is an illustration of the trivia about which so many students complain. Furthermore, such figures are seldom as precise as they appear.

(5) *Be certain the stem is accurate.* Consider:

Why did Germany want war in 1914?
1. She was following an imperialistic policy.
2. She had a long-standing grudge against Serbia.
3. She wanted to try out new weapons.
4. France and Russia hemmed her in.

Who is in any position to say that Germany wanted war? Such inexactitudes may strengthen misinformation on the part of students.

(6) *Adapt the level of difficulty of the item to the group and to the purpose for which the item is intended.* Consider:

If a tree is growing in a climate where rainfall is heavy, are large leaves an advantage or a disadvantage?
1. An advantage, because the area for photosynthesis and transpiration is increased.
2. An advantage, because large leaves protect the tree during heavy rainfall.
3. A disadvantage, because large leaves give too much shade.
4. A disadvantage, because large leaves absorb too much moisture from the air.

The above item illustrates an increased level of difficulty because it requires knowledge of both the answer and an explanation for it.

(7) *Omit clues to the correct response.* Items that contain clues or cues are not measuring what the instructor intended. Including clues is perhaps the most frequent error made in multiple-choice tests. In the following item it is necessary only to know that "exert" is commonly used with "pressure":

What does an enclosed fluid exert on the walls of its container?
1. Energy
2. Friction
3. Pressure
4. Work

In the next item the stem calls for a plural answer, which occurs only in 4.

Among the causes of the Civil War were:
1. Southern jealousy of northern prosperity.
2. Southern anger at interference with the foreign slave trade.

3. Northern opposition to bringing in California as a slave state.
4. Differing views on the tariff and Constitution.

In the next item the correct answer has been stated more precisely and at greater length than the others. Students catch on quickly to such a clue.

Why were the Republicans ready to go to war with England in 1812?
1. They wished to honor our alliance with France.
2. They wanted additional territory for agricultural expansion and felt that such a war might afford a good opportunity to annex Canada.
3. They were opposed to Washington's policy of neutrality.
4. They represented commercial interests which favored war.

In the next item there are common elements in the stem and in the answer:

What led to the formation of the States Rights Party?
1. The level of federal taxation
2. The demand of states for the right to make their own laws
3. The industrialization of the South
4. The corruption of many city governments

Finally, such specific clues as "all," "always," "certainly," and "never" are to be avoided—they are clues to incorrect answers. Moreover, scholars are leery of absolutes and probably should encourage students to be.

(8) *Do not use a negatively stated item stem.* Experience has shown that these tend to confuse students, yet some items contain two and three negatives and seem like intricate verbal puzzles.

Which of these is *not* one of the purposes of Russia in consolidating the Communist party organization throughout Eastern Europe?
1. To balance the influence of the western democracies
2. To bolster her economic position
3. To improve Russian-American relations
4. To improve her political bargaining position

Which of these is *not* true of a virus?
1. It is composed of very large living cells.
2. It can reproduce itself.
3. It can live only in plants and animal cells.
4. It can cause disease.

(9) *Be certain that the correct answer is one on which competent critics agree.* Consider:

What is the chief difference in research work between colleges and industrial firms?
1. Colleges do much research, industrial firms little.
2. Colleges are more concerned with basic research, industrial firms with applications.
3. Colleges lack the well-equipped laboratories which industrial firms maintain.
4. Colleges publish results, while industrial firms keep their findings secret.

Competent authorities could not agree upon the best response to the above. If this type of item is to be used, a qualification should be offered in the stem, such as, "According to_____, the chief difference...."

(10) *Avoid answer alternatives that overlap or include each other:*

> What percent of the total [property] loss due to hail is the loss of growing crops?
> 1. Less than 20 percent
> 2. Less than 30 percent
> 3. More than 50 percent
> 4. More than 95 percent

If 1 is correct, then 2 is also correct; and if 4 is correct, then 3 is correct.

This discussion is not intended to suggest that test questions for college students should be simple or tests easy. For the most part, the examples emphasize item clarity; they do not deal with what should be measured—factual information, concepts, appreciation, and so on. Many authorities believe that multiple-choice items, if constructed with great care, can measure conceptual knowledge, ability to generalize, and so forth. The way to prepare such items is to be clear about one's own objectives of instruction and to enlist the assistance of one's colleagues in judging whether a particular item measures what is intended.

Essay Questions

For a variety of reasons, essay questions or items require less preparation time than multiple-choice ones; on the other hand, the essay type requires much more time to score. We estimate, however, that for classes numbering around 35 students the instructor would invest about equal time for properly prepared multiple-choice tests and for properly scored essay ones. Faculty time, however, is not the sole criterion for deciding between the two types of test. The essay question, permitting freedom of response, can test how students approach a problem, what information they think is important, and what conclusions they reach. Debates continue over other qualities or abilities that essay questions are purported to measure (for a research review, see Yeasmeen and Barker).

Whatever the merits and faults of essay questions, they afford students an opportunity to express themselves in their own words, as Stalnaker, among many others, has emphasized. Essay questions compel students to think about a topic, decide what to say about it and how to say it, and do the writing. These are important abilities in an educated person, and many faculty members are convinced that the development of these abilities has been deterred by the excessive use of objective tests. At the very least, essay questions give students an incentive to write.

Most of the principles for promoting multiple-choice item clarity apply equally to essay questions. The application of several additional principles will increase the chances of attaining scoring consistency (or reliability).

(1) *Limit the scope of the question.* There is simply no way of scoring fairly such broad questions as "Discuss Shakespeare's tragedies" or "Analyze the energy crisis." Moreover, students must guess which replies will please the instructor—they must "psych out the prof."

Restrictions of the scope may vary, of course; limits may be imposed by calling for brevity and conciseness, insisting upon only a few sentences, or even specifying the space to be used. Questions may be structured in other ways—by asking students to compare, contrast, discriminate, note limitations, draw inferences, state conclusions tersely, and so on.

(2) *Avoid items or questions that are based on personal feelings.* Educators are in no position to measure or quantify students' feelings about any issue. Such questions as, What does modern art mean to you?, Do you relate to the writings of e.e. cummings?, and How do you feel about Truman as a President? promote "psyching out the prof." If the answers are honest, there are no standards by which they can be quantified. To many people modern art means nothing; others cannot abide e.e. cummings, and well-informed persons differ about Truman. Where the affective domain is concerned, unfair and improper judgments are more likely to be rendered on official records when students' feelings and opinions do not agree with those of their instructors.

(3) *Be certain that an adequate answer can be given in the time allowed.* It is amazing how often this simple rule is violated, even by those who know from personal experience how difficult it is to organize thoughts and present them coherently. Again, the issue is what is being measured; the quickest student is not necessarily the best one in all respects.

(4) *Use the following procedures for scoring essay items,* bearing in mind that the subject is measurement and the goal of measurement is objectivity:

- Minimize, as far as possible, cues that will identify the owners of the papers; at the very least, remove the names. It is all too easy to allow extraneous knowledge about a student to influence the marking of his or her paper. Such precautions should help to assure minority students that the marking process is free of discrimination.
- Write out an ideal answer ahead of time and ask a colleague to do likewise; combine the two into a standard with which students' replies can be compared.
- Score each item on a point scale without reference to a passing grade (assigning a grade is evaluation, not measurement); that is, determine prior to scoring that item 1 can earn 10 points, for example, item 2 is worth 20 points, and so on. Total all the points and then assign a letter grade.

Test Construction

Most tests are composed of several items or questions that are put together for some specific purpose, measuring students' ability to translate a foreign language, for example. Questions on a given test constitute a sample of all the questions that could be asked. There are no hard and fast rules that will produce a representative sample of questions, but there are guidelines that will increase the chances of a fair distribution:

• *Have the items reflect important objectives you have attempted to promote.* This guideline is difficult to elaborate because goals or objectives will vary from course to course. The trick is to aim for an unbiased sample of questions. In Case 1 (page 13), the test was biased in that the majority of questions were on Kant although goals of the course were not limited to understanding that gentleman. Perhaps the simplest way to avoid an unduly biased sample of questions on a particular test is to have a colleague criticize it before the test is given.

• *Generally speaking, the greater the number of items in a test, the more representative the sample.* This is one of the arguments in favor of multiple-choice items. In a given period of time, more multiple-choice than essay questions can be answered.

• *Allow ample time for all students to respond to all the questions.* The experience of colleagues about the optimum length of a test for a given time period will be helpful.

Regardless of the care with which tests are constructed, there will be errors just as there are errors in all measurement. In physical measurement, the errors stem from at least two sources, defects in the measuring instrument and perceptual distortions associated with the person taking the reading. For educational measurement there is an additional source—the person being measured. The performance of anyone tends to fluctuate from day to day for a variety of reasons. The goal, then, is to minimize errors in measurement.

Errors in Test Construction

Regardless of the care with which tests are constructed, there will be errors just as there are errors in all measurement. In physical measurement, the errors stem from at least two sources, defects in the measuring instrument and perceptual distortions associated with the person taking the reading. For educational measurement there is an additional source—the person being measured. The performance of anyone tends to fluctuate from day to day for a variety of reasons. The goal, then, is to minimize errors in measurement.

Regarding the instrument or test, clearly written items and a representative sampling of material will decrease errors and increase reliability (i.e., consistency or stability). Also, generally speaking, the longer a test, the greater its reliability. Multiple-choice tests tend to be more reliable than essay ones because more questions can be answered in a specified period of time.

As for reliability of marking, properly prepared multiple-choice

tests are the least subject to error. Scoring of essay tests tends toward unreliability or inconsistency. Two or more instructors are likely to arrive at different scores, and the same instructor may arrive at different scores at different times. The reliability of a given test can be determined statistically, and for large introductory courses such determinations are very much in order. Ebel (1972) has estimated the average reliability of college tests to be .45, a coefficient that reflects unreliability, inconsistency, and imprecision. (A perfect coefficient of reliability is 1.00.)

The third source of error, the person being measured, encompasses the day-to-day personal variations we all experience and the conditions under which a test is administered. These can help or hinder performance. For the purpose of averaging out day-to-day variations, conducting several tests during a course will tend to yield more reliable or consistent measures than giving a single one. Although we do not recommend a specific number of tests, it is clear that giving only one test for an entire course is likely to be unreliable. As for physical conditions in a room used for a test, inadequate ventilation, uncomfortable temperatures, poor lighting, or excessive crowding will tend to cause inaccurate measures; in Case 12 (page 16), cheating resulted in inaccurate scores. Poor testing conditions are inexcusable.

Rhodes has summarized the meaning of errors of measurement:

> It is assumed that for each test a student takes, there is a true score he should make that may differ from the score he actually achieves. The true score would be free of the accidental error caused by factors such as the questions selected for the test, how the student feels on the day of the test, the temperature of the testing room, and so on. Theoretically, if a student took an infinite number of equivalent editions of a test, the scores he obtained would vary somewhat but would cluster around an average, or true score. The score a student actually obtains on any given occasion is, then, an approximation of this true score and should be thought of as representing an interval, or obtained-score range, the limits of which are determined by use of the standard error of measurement.

Finally, in this respect, there is a statistical formula for calculating the standard error of measurement that is useful for large classes.

A test can be very reliable, yielding precise and accurate scores, but really not measure anything of importance. Such a test is of course invalid. While there are several concepts of validity (for detailed discussions, see Ebel, 1972), only two need be of direct concern—content validity and predictive validity.

Content validity means that a test measures what it is supposed to measure, for example, critical thinking about economics or problem solving in calculus. Well-formulated objectives for a course are the first prerequisite for attaining content validity of test items. The second requirement is the advice of one's colleagues.

The multiple-choice items in the following table presumably have content validity.

Multiple-Choice Items Intended to Test
Various Aspects of Achievement*
Understanding of Terminology or Vocabulary

The term "fringe benefits" has been used frequently in recent years in connection with labor contracts. What does the term mean?

1. Incentive payments for above-average output
2. Rights of employees to draw overtime pay at higher rates
3. Rights of employers to share in the profits from inventions of their employees
4. *Such considerations as paid vacations, retirement plans, and health insurance*

What is the technical definition of the term "production"?

1. Any natural process producing food or other raw materials
2. *The creation of economic values*
3. The manufacture of finished products
4. The operation of a profit-making enterprise

Knowledge of Fact and Principle or Generalizations

What principle is utilized in radar?

1. Faint electronic radiations of far-off objects can be detected by supersensitive receivers.
2. *High-frequency radio waves are reflected by distant objects.*
3. All objects emit infrared rays, even in darkness.
4. High-frequency radio waves are not transmitted alike by all substances.

The most frequent source of conflict between the western and eastern parts of the United States during the course of the nineteenth century was:

1. *The issue of currency inflation*
2. The regulation of monopolies
3. Internal improvements
4. Isolationism vs. internationalism
5. Immigration

Ability to Explain or Understanding of Relationships

If a piece of lead suspended from one arm of a beam balance is balanced with a piece of wood suspended from the other arm, why is the balance lost if the system is placed in a vacuum?

1. The mass of the wood exceeds the mass of the lead.
2. The air exerts a greater buoyant force on the lead than on the wood.
3. The attraction of gravity is greater for the lead than for the wood when both are in a vacuum.
4. *The wood displaces more air than the lead.*

Should merchants and middlemen be classified as producers or nonproducers? Why?

1. As nonproducers, because they make their living off producers and consumers
2. As producers, because they are regulators and determiners of price
3. *As producers, because they aid in the distribution of goods and bring producer and consumer together*
4. As producers, because they assist in the circulation of money

Ability to Calculate or Numerical Problems

If the radius of the earth were increased by three feet, its circumference at the equator would be increased by about how much?

1. 9 feet
2. 12 feet
3. *19 feet*
4. 28 feet

What is the standard deviation of this set of five measures—1,2,3,4,5?

1. 1
2. $\sqrt{2}$
3. 9
4. $\sqrt{10}$
5. None of these

*Adapted from Exhibit 5.2, Robert L. Ebel, **Essentials of Educational Measurement**, © 1972, pp. 111-113. Reprinted with permission of Prentice-Hall, Inc.

Ability to Predict or What is Likely to Happen Under Specified Conditions?

If an electric refrigerator is operated with the door open in a perfectly insulated sealed room, what will happen to the temperature of the room?

1. *It will rise slowly.*
2. It will remain constant.
3. It will drop slowly.
4. It will drop rapidly.

What would happen if the terminals of an ordinary household light bulb were connected to the terminals of an automobile storage battery?

1. The bulb would light to its natural brilliance.
2. *The bulb would not glow, though some current would flow through it.*
3. The bulb would explode.
4. The battery would go dead in a few minutes.

Ability to Recommend Specific Appropriate Action

Which of these practices would probably contribute least to reliable grades from essay examinations?

1. *Weighting the items so that the student receives more credit for answering correctly more difficult items.*
2. Advance preparation by the rater of a correct answer to each question.
3. Correction of one question at a time through all papers.
4. Concealment of student names from the rater.

"None of these" is an appropriate response for a multiple-choice test item in cases where:

1. The number of possible responses is limited to two or three.
2. *The responses provide absolutely correct or incorrect answers.*
3. A large variety of possible responses might be given.
4. Guessing is apt to be a serious problem.

Ability to Make an Evaluative Judgment

Which one of the following sentences is most appropriately worded for inclusion in an impartial report resulting from an investigation of a wage policy in a certain locality?

1. The wages of the working people are fixed by the one businessman who is the only large employer in the locality.
2. Since one employer provides a livelihood for the entire population in the locality, he properly determines the wage policy for the locality.
3. Since one employer controls the labor market in the locality, his policy may not be challenged.
4. *In this locality, where there is only one large employer of labor, the wage policy of this employer is really the wage policy of the locality.*

Which of the following quotations has most of the characteristics of conventional poetry?

1. "I never saw a purple cow;
   I never hope to see one."
2. *"Announced by all the trumpets of the sky*
   Arrives the snow and blasts his ramparts high."
3. "Thou art blind and confined
   While I am free for I can see."
4. "In purple prose his passions he betrayed
   For verse was difficult.
   Here he never strayed."

The predictive validity of college tests is low. That is, scores derived from them do not predict future performance very well. For a better understanding of predictive validity, considerable research

is needed to determine what magnitude of difference between scores is significant. Often student X with a score of 91 will receive an A, while student Y with 89 will receive a B. For GPA purposes on most campuses these translate into 4.00 and 3.00, respectively. It is assumed that student X can and will out-perform student Y, but the evidence that this is true is tenuous. How large must the difference be between the two—1, 5, 10, 20 points or more—before the predictive assumption is substantiated?

Why is precision emphasized? Because CPAs are used in an exceptionally precise manner, as when arbitrary cut-off scores are set. A 3.50 may entitle a student to further consideration for admission to a program, while a 3.49 results in categorical rejection. Under these circumstances the least that can be striven for is accuracy in measurement.

4

Grading

A comprehensive evaluation of student performance should provide guidance for academic improvement, but students too often receive scant critical commentary on their progress. Letter grading, the most commonly accepted form of evaluation, is particularly susceptible to the charge of insufficient feedback to the student. A more fundamental grasp of the options for academic measurement is the most direct route to improved grading.

EVALUATIONS SHOULD MEAN PROVIDING A GREAT DEAL OF information to students about their academic performance—strengths, deficiencies and corrective steps to be taken, relative standing, and other pertinent details. Blum has observed in this connection: "It is no secret that students often receive little critical commentary on their papers and examinations. The result is that the prospects for academic improvement are diminished...."

There is this paucity of detailed help for students because evaluation now tends to mean the assigning of letter symbols for record-keeping purposes. The subject of grading is laden with prejudices, dogmas, and unfounded opinions, and for many years it has tended to provoke very unscholarly pronouncements. It is not a new dilemma. In 1890, a Virginia institution had a six-point grading scale—optimus, melior, bonus, malus, pejor, and pessimus. Because the president thought too many mediocre students received the grade of optimus, the scale was changed to a three-point one—distinguished, approved, and disapproved. Soon, however, the president was discontented again, for "some bad scholars were approved, and good scholars were all distinguished" (Cureton).

The purpose in mentioning letter grading is to stimulate scholarly attention to the subject. Such attention is imperative if progress is to be made. Our discussion of the unresolved issues associated with the assigning of grades is followed by some tentative suggestions for improvement.

One reason some of the problems here are not yet being resolved is the fact that several assumptions have not been examined except by a few specialists. Another is the widespread and comfortable belief inside and outside academe that letter grades have considerable

predictive validity. In truth, they do not. McClella...d has summarized data about the predictive value of grades:

> Researchers have in fact had great difficulty in demonstrating that grades in school are related to any other behavior of importance.... It seems so self-evident to educators that those who do well in their classes must go on to do better in life that they systematically have disregarded evidence to the contrary that has been accumulating for some time.

In a recent survey of studies about grades, Warren found that about half of approximately 200 articles, papers, and reports that appeared between 1965 and 1970 dealt with the form of grades (A,B, C,D,F, P,F, etc.) and with grades as predictive measures. The other half were concerned with a variety of aspects, such as presumed advantages and disadvantages. Warren concluded: "These reports, in spite of their variety, leave large gaps in our knowledge about grades and grading.... These results do not constitute an impressive advance in knowledge about an important, ubiquitous process in higher education...."

There 's this paucity of detailed help for students because evaluation now tends to mean the assigning of letter symbols for record-keeping purposes. The subject of grading is laden with prejudices, dogmas, and unfounded opinions, and for many years it has tended to provoke very unscholarly pronouncements.

Problems

Single course grades are used to compare students within an institution and across institutions. If measurements are the basis of a comparison, no two of anything, let alone the learning of two people, can be compared unless the same instrument is used for both measurements. Woe be to the cabinetmaker who tries to assemble pieces of rare and exotic wood some of which he has measured with a giveaway yardstick and others with a finely calibrated meter stick. For physical measurements, of course, there are many agreed-upon scales or units—inch, yard, mile, ounce, pound, ton. Each of these can be determined precisely so that two or more measurements in the same units tend to have quite exact meaning. A pound on the West Coast has the same meaning as a pound on the East Coast. Perhaps the basic problem in grading students for purposes of comparison is the absence of any such agreed-upon measurements.

A second problem is inherent in the uncritical acceptance of norm-referenced grading, or what students refer to as "grading on

the curve." This may have come into extensive use because of the need to compensate for the lack of a measuring unit. At any rate, norm-referenced grading derives from the mythical "normal curve of distribution" or bell-shaped curve. Its pervasive and often distorted applications have created an illusion of the existence of a standard by which students can be compared equitably, first by the professor who assigns the symbol and then by all others who see it. In fact, the "normal curve" is nothing more than a mathematical ideal or model. Moreover, according to Lindquist, there is an erroneous belief that mental ability test data have been shown to form the bell-shaped curve. The overlooked fallacy is that many standardized tests are constructed deliberately so that the scores will yield such a curve; in some cases foxy statisticians manipulate the scores.

The potency of the false standard is illustrated by this episode (Dressel):

> In one university, the decision was made to section engineering students in calculus on the basis of previous grades. One professor, not knowing this, was assigned a group of students in integral calculus who had received A's in all preceding mathematics courses. Although recognizing that this was an unusually good group...on the first examination, he ended up with the usual distribution of grades from A to F. The reaction of the students forced him to reconsider.... The grades at the end of the term showed 40 percent A's, 50 percent B's, and 10 percent C's. Knowing the caliber of the students, the professor still could not bring himself to report a distribution of grades in which almost every student would be given an A.

This professor thought he had firm reference points for setting cutoff scores for each grade.

It is bad enough when a lone professor grades on the curve for a single class of highly capable students. It is even worse when a gifted student body is judged in this manner. Reed College has established grade guidelines for all faculty to follow (Levine and Weingart). For freshmen the distribution is supposed to be A, 15 percent; B, 35 percent; C, 40 percent; D, 10 percent. For the remaining three categories of students, the recommended distribution is A, 15 percent; B, 45 percent; C, 35 percent; D, 5 percent. Needless to say, such grading can cause talented students to encounter difficulties in being admitted to graduate and professional schools. In the final analysis, grading on the curve means statistical relativism; students are rank-ordered from high to low.

Grade point averages are also used to compare students within an institution and across institutions. Basic errors in testing and grading are compounded by the numerous ways in which GPAs are computed at different institutions. In one survey of these practices (Collins and Nickel) from a sample composed of 650 public and private two- and four-year institutions in the 50 states and the District of Columbia, with 448 schools responding, great variation was found (see table on next page).

The survey revealed that in some schools such grades as Incompletes immediately become F's for calculation purposes, while in others more than an entire term can elapse before such academic capital punishment is applied. As one example of "sudden death," during experimental investigation of instruction at the University of

Texas at Austin (Stice) it was necessary for students to receive Incompletes if they desired. During one term 26 percent did so. None of the investigators knew of the policy that I's became F's for GPA purposes nor did several staff members in the registrar's office. Several good students lost scholarships and others failed to receive invitations to honor societies. More than likely the calculation practices are not specified on very many transcripts.

The assumption that single grades have common reference points has been made about GPAs, too. Who knows what sorts of tests are behind the grades or the standards by which the grades were derived? If anything, GPA statistics as they are presently employed tend to be meaningless—despite what most academicians and others think.

Our numerous deliberations about grading led repeatedly back to several basic facts: (1) Unidimensional symbols report multidimensional phenomena. A given grade can reflect level of knowledge, at-

Practice	Number of Institutions Indicating This Is Present Practice
All grades received in all courses taken at any institution are used in computing the overall grade point average.	159
Only grades in courses which count for the degree are used in computing the GPA.	43
Only grades in courses taken in the institution doing the computing are used in computing the GPA.	246
When a course is repeated, all grades (two or more) are used when computing the GPA.	136
When a course is repeated, only the last grade received is used when computing the GPA.	266

titudes, procrastination, interest or lack of it, and other factors. The lone symbol specifies none of these things. Perhaps each professor assumes that every other interpreter will see in the lone symbol all of the nuances he or she intended. (2) The symbol, by itself, reveals nothing about the quality of the test or tests through which it has been derived.

Suggestions

An emerging model of grading is called criterion-referenced. Its basic feature is the concept of mastery. If anything, criterion-referenced grading requires more complete statements of objectives than does norm-referenced grading. Tests are designed, then, to deter-

mine whether a student has or has not attained these objectives. The concept of criterion-referenced grading has been used especially in the Keller Plan (see Ruskin and Hess) and in contract grading. (While this approach appears to be more and more common, there is little about it in the literature.) There are several excellent references for criterion-referenced grading—Popham, Carver, and Angoff.

Criterion-referenced grading is used in the emerging competency-based curricula. For a digest of its important features in this context (as well as answers to questions that are being asked such as, What is competence? and How does the faculty role change in a competence curriculum?), see the report by the Southern Regional Education Board.

This method of grading certainly has its place, especially in professional curricula. When it is used for a given course, a notation should be made on the transcript to facilitate interpretation.

Finally, there is the import for grading of the basic theme of this volume—*improved testing or measurement is the fundamental route to improved grading*. There are no substitutes for clarity about what one is trying to accomplish in instruction and very careful efforts to find out what students have achieved.

Our numerous deliberations about grading led repeatedly back to several basic facts: [1] Unidimensional symbols report multidimensional phenomena. A given grade can reflect level of knowledge, attitudes, procrastination, interest or lack of it, and other factors. The lone symbol specifies none of these things. Perhaps each professor assumes that every other interpreter will see in the lone symbol all of the nuances he or she intended. [2] The symbol, by itself, reveals nothing about the quality of the test or tests through which it has been derived.

Etzioni recently suggested that what is needed is open discussion by departments leading to agreement about grading standards, but this would be insufficient. Once again the tip of the iceberg would be considered while its submerged body would be ignored. A better solution would be open discussions by departments about all facets of testing. A professor can no longer go it alone in certifying students for society.

5

Lone Efforts Are Not Enough

Growing external pressures are forcing faculty to take a fresh look at student evaluation. The new consumerism, recent legal decisions, and far-reaching social criticism will no longer leave matters of grading and testing to the private academic preserve. The use of external examiners and the establishment of effective campus grievance arrangements are only two of the ways recommended to improve an increasingly nettlesome issue in academic life.

IF ASSESSMENT IS NOT IMPROVED FROM INSIDE THE PROFESsion, then it most surely will be put under pressure from the outside. Traditionally faculty members have enjoyed almost complete autonomy in their teaching performance. Until recently the courts had tended to avoid the academic bastions. But now they are beginning to intervene, and some observers believe such intervention will soon accelerate. This has resulted from several trends: an increased sophistication of students, a new regard for higher education as a social necessity and an individual right, the expansion of civil rights protections by public authority, and—perhaps most important—the new age of majority.

The Courts Intervene

One instance of recent court intervention dealt with a lone grade (*State Ex Rel. Bartlett v. Pantzer*). A political science student graduated from the University of Chicago in June 1971 with a Bachelor of Arts degree. During the spring quarter of his senior year he had enrolled in a graduate accounting course to fulfill an admission requirement of the law school of the University of Montana, where he was seeking admission in September 1971. The law school had informed the student that the requirement would be fulfilled if he received a satisfactory grade.

The student received a D in the course, whereupon he was advised by the law school that he would not be admitted because the grade was not a satisfactory one. Testimony in court revealed that

colleges and universities regarded a grade of D as "acceptable," but not "satisfactory." The Supreme Court of Montana was unable to discern the exquisite difference and directed the law school to admit the student.

More court intervention in matters of academic measurement seems likely in the not-too-distant future. The United States Supreme Court made a momentous decision in the *Griggs v. Duke Power Company* case and may have set a precedent for drastically altered interpretations of higher education test scores and grade point averages. The company was found to have discriminated racially by requiring, for an employee to be promoted from laborer to coal handler, either the possession of a high school diploma or the passing of two standardized tests. In rendering its decision, the court ruled: "Nothing in the act (Civil Rights Act, 1964, Title VII) precludes the use of testing or measuring procedures; obviously they are useful. What Congress has forbidden is giving these devices and mechanisms controlling force unless they are demonstrably a reasonable measure of job performance."

Suits have been instituted already in several states charging that bar examinations discriminate unfairly against minority groups.

If assessment is not improved from inside the profession, then it most surely will be put under pressure from the outside. Traditionally faculty members have enjoyed almost complete autonomy in their teaching performance. Until recently the courts had tended to avoid the academic bastions. But now they are beginning to intervene, and some observers believe such intervention will soon accelerate.

The fundamental issue is the predictive validity of such tests for all who take them. It could well be that these assaults upon bar exams are a prelude to assaults on many other licensing examinations, because they, too, are job related. Since higher education in its testing activities is engaged more in credentialing or rank-ordering students than in assessing learning, it is not too difficult to foresee grade point averages being ruled job-related by the courts. (Today a student may be refused admission to a professional school because of a GPA a few hundredths of a point below some arbitrary cut-off score.) Many ramifications of the *Duke Power Company* decision and its innumerable complexities have been examined meticulously and thoughtfully by Huff.

Of more direct portent for the future may be the dissenting opinion of former Justice William O. Douglas in *DeFunis v. Odegaard* (Fields). Justice Douglas was especially critical of scores derived from the Law School Admissions Test and of grade point averages and the fact that they had dominated the selection process. He argued that law schools are not bound to admit students according to mechanical criteria because such criteria often conceal important

abilities. Justice Douglas was most persuasive in his plea for more thorough assessment of individual attributes than test scores provide. For example, he maintained that a person who pulls himself from the ghetto via a community college has demonstrated a quality of perseverance and thereby has more promise for the study of law than a rich graduate of Harvard. The poorer applicant should be admitted, said Douglas, because he had shown special potential in contrast to the Harvard graduate who may have taken less advantage of the vastly superior opportunities afforded him.

It is too soon to know the full impact of the so-called Buckley Amendment that gives students access to their test papers and other official records, but scores of students may avail themselves of the access and be so overwhelmed that they will demand careful and honest explanations for selected test scores and grades. This provision of law may give them a basis for court action to enforce their demands. Quite obviously, poor tests and unfair grades are features of instruction that are under the direct control of each individual faculty member. Just how could a student's "improper spirit toward the subject matter" (Case 9, page 15) be documented or substantiated in court?

Unless professors individually and collectively begin to make drastic improvements in testing and grading practices, there will be intrusions on their autonomy from without in several forms. There even appears to be a possibility of compulsory state or nationwide standardized tests of academic achievement. Academic freedom is imperative and must be preserved, but the professoriate cannot avoid its own responsibilities. Grading policies and practices in most undergraduate courses do not bear any relation to inviolable academic freedom.

What does all this mean? Unless professors individually and collectively begin to make drastic improvements in testing and grading practices, there will be intrusions on their autonomy from without in several forms. There even appears to be a possibility of compulsory state or nationwide standardized tests of academic achievement. Academic freedom is imperative and must be preserved, but the professoriate cannot avoid its own responsibilities. Grading policies and practices in most undergraduate courses do not bear any relation to inviolable academic freedom.

How, then, can the process be improved? Classroom tests can be improved by faculty members learning more about measurement and obtaining the assistance of their colleagues. At least three additional reforms must be implemented to improve the test product and demonstrate the professoriate's willingness to put its house in order.

Visiting Examiners

It is a deeply ingrained belief throughout American higher education that instructing and examining are inseparable. The instructor is supposedly the person best able to judge the work of his or her student.

There has been at least one historical challenge to this assumption (Coulter). In 1811 the three trustees of the University of Georgia were named as visitors and urged, along with other distinguished men of the state, to attend examinations of seniors because: " 'The test of the pudding is the taste thereof' is a saw honored with age and truth. Examination times were tasting times and this tasting should be done by more than the cooks only." By 1825 the examinations for juniors were being attended by any person who desired to attend.

A modern and refined counterpart to this practice of some 150 years ago is the visiting examiners tradition for the Honors Program of Swarthmore College (Swarthmore College Faculty, 1941), which began in the early 1920s, continues to flourish today, and is widely acclaimed by faculty, students, and alumni.

Around 40 percent of juniors and seniors elect to take honors work. Normally this means that a student studies six subjects during the last two years. The work is pursued independently or in small seminars. At the end of the senior year the student is subjected to a three-hour written examination in each subject. These exams are prepared and evaluated by faculty members from other institutions. In the oral examinations that follow, there is no rigid pattern; they are conducted in a variety of ways. But the judgment of the visitor carries the most weight.

A recent evaluation of the program (Swarthmore College, 1967) describes the rationale and the benefits succinctly:

> Many external examiners...think the system works well, and the examiners' evaluations of students are generally consistent with the faculty's. Many graduates of honors have said (in the alumni questionnaire), as have many faculty, that the system helps to create an atmosphere of faculty-student collaboration.... These are now conventional statements; but we are inclined to agree with them. The colleagueship and the intellectual checks provided by external examiners are widely felt to be valuable for both students and the faculty; many of the latter, especially, set high store by it....

On all too many campuses faculty and students are two factions warring over learning. The faculty are so dedicated to the exercise of their selective function, they cannot see teaching-learning as a collaborative endeavor, whereas at Swarthmore apparently faculty members and students work together to meet and impress a sort of common foe, the visiting examiner. Thus one reason for more extensive use of this type of program is that it serves the cause of learning for the individual students who participate.

A second reason for having visiting examiners on many campuses is that their presence should broaden the perspectives of faculties about the art and techniques of teaching. While the various faculty organizations help keep the professoriate abreast of disciplinary developments, many pay little direct attention to good teaching. With-

out sufficient stimulation it is very easy to become smug, myopic, and provincial. If, over a substantial period of time, too many students performed poorly, the visiting examiners would be in a position to ask some penetrating questions of the home faculty. Help by colleagues from other institutions is more useful and more palatable than interference from those outside academic life.

Testing Specialists

Another challenge to the notion that teaching and testing are inseparable came during the early 1930s at the University of Chicago with the creation of the Board of Examinations (Bloom). The faculty were

---

Recently, perhaps partly as a result of the joint statement, grievance procedures have been made formal in some institutions and often include a specially appointed committee, which in some cases is given the authority to overrule a faculty member and change a grade. For example, at California State University, Los Angeles, if a grade grievance is not resolved at the departmental level, the student may appeal to the dean of that school who, in turn, refers the matter to a special committee. The dean, after consultation with the committee, may authorize a change of grade. If for any reason a student believes the problem has not been resolved fairly, he or she may submit a signed statement to the standing student grievance committee, which may refer the issue to one of several other committees, any one of which may recommend a grade change to the appropriate dean, whereupon the change is made in the permanent records.

---

concerned primarily with having students assume responsibility for their own learning. Degree requirements were set in terms of comprehensive examinations, and as a result students could make individual decisions about the speed with which they would attain their degrees as well as about their study methods and class attendance.

Since the comprehensive examinations were the sole basis for meeting graduation requirements, they had to be excellent measures of academic achievement. In consultation with faculties, a corps of test specialists constructed the exams, scored them, and assigned grades. The faculty believed that an ideal teacher-student relationship—one which promoted an optimum of learning—was impossible when the teacher also served as judge and jury. The success of the project was revealed, in part, by the high test relia-

bility coefficients that were obtained. These ranged almost without exception between .90 and .95.

Several forces combined during the early 1950s to eliminate this extreme departure from traditional testing and grading practices. In the meantime, several campuses have established offices that serve instructors on a voluntary or request basis. One example is the Evaluation and Examination Service of the University of Iowa (Whitney). The service staff consults with individual faculty members or departments on techniques of test construction and improvement, test and item analysis, and methods of grade assignment. In addition, course examinations are duplicated, scored, and analyzed. The service keeps the faculty and others informed periodically by means of memos and technical bulletins. A current memo is entitled, "Should I Take the Graduate Record (GRE) Again?" Recent bulletins discussed "Improving Essay Questions." There are two professional members of the staff; about 40 percent of a faculty of 700 use the service. Comparable agencies should be available to faculties on all campuses.

In *Change*'s first faculty policy paper on professional development, the authors, in a chapter entitled "Evaluation for What?", suggest the ideal of the separation of teacher from evaluator: "A developmental approach to education calls for a new kind of detachment: for students the detachment of the process of learning from the certification of competence; and for teachers, detachment of efforts to improve teaching from official assessments of performance" (Group for Human Development in Higher Education).

Academic Grievances Committees

Tradition has it that if a student feels a grade is an improper one, he or she may seek redress by consulting the individual faculty member. If satisfaction is not received, the student has had the right to consult with other individuals—department heads, deans, and even the president or chancellor. For the most part the arrangements have been informal and final authority to change or not change the grade has rested with the faculty member.

In 1967 several important organizations* issued a Joint Statement on Rights and Freedoms of Students. The statement included this right: "*Protection Against Improper Academic Evaluation*—Students should have protection through orderly procedures against prejudiced or capricious academic evaluation. At the same time, they are responsible for maintaining standards of academic performance established for each course in which they are enrolled."

Recently, perhaps partly as a result of the joint statement, grievance procedures have been made formal in some institutions and often include a specially appointed committee, which in some cases is given the authority to overrule a faculty member and change a grade. For example, at California State University, Los Angeles, if a grade grievance is not resolved at the departmental level, the student may appeal to the dean of that school who, in turn, refers the

*American Association of University Professors, U.S. National Student Association, Association of American Colleges, National Association of Student Personnel Administrators, and National Association of Women Deans and Counselors.

matter to a special committee. The dean, after consultation with the committee, may authorize a change of grade. If for any reason a student believes the problem has not been resolved fairly, he or she may submit a signed statement to the standing student grievance committee, which may refer the issue to one of several other committees, any one of which may recommend a grade change to the appropriate dean, whereupon the change is made in the permanent records.

At Western Michigan University, the arrangements are less complicated. If a student is dissatisfied following informal consultation within the department, he or she may see an administrator, who may decide the grievance is unwarranted or there is sufficient evidence for the case to be considered by a committee on academic fairness, either the graduate or the undergraduate committee. The undergraduate committee consists of three faculty members, three undergraduates, and a nonvoting chairperson. If the committee decides to recommend a change of grade, the faculty member is informed first so that he or she may make the change. If the faculty member prefers not to do so, the committee then makes the change by notifying the dean of records and admissions.

At Pomona College, the procedures are simple and straightforward. If, after the usual informal hearings, the disputants are still disgruntled, the dean appoints a small ad hoc committee of faculty from the department of the instructor or from a related department. "The decision of the hearing committee on the disputed grade shall be final."

There are formal hearing procedures in other institutions, but in these the final judge—whether a committee, a dean, or a chancellor—has no power to change a grade. Appeals for fairness can be addressed to the faculty member, but not a decision that a grade must be changed. After going to elaborate lengths to ensure academic rights for students, Michigan State University (1969) persistently maintains the traditional stance that the instructor is the only person who can assign a grade. In most instances instructors are cooperative, but nothing further can be done if they stubbornly defy the grievance committee, according to an official.

We recommend that formal arrangements be established for reconciling testing and grading grievances and that a final judge other than the instructor have the authority to change a grade. This recommendation is made for these reasons:

(1) Cases such as some of those mentioned in the first chapter reflect almost unbelievable examples of faculty arbitrariness and capriciousness. Students should be able to fight back against such unfairness, and with the balance of power on their side. This presumes our basic system of justice, which is designed to protect the rights of the weak individual who is being persecuted by strong external authorities.

(2) The mere existence of such appeal arrangements should help decrease testing and grading offenses.

(3) Correction by one's peers is both more palatable and more effective than intrusion by outside forces.

6

For Further Reading

For a more comprehensive understanding of testing and evaluation, faculty have access to a number of excellent source documents. Here are some of the best.

References

Adkins, Dorothy Wood. *Test Construction: Development and Interpretation of Achievement Tests.* Columbus: Charles E. Merrill, 1960.

Angoff, William H. "Criterion-Referencing, Norm-Referencing, and the SAT." *College Board Review*, no. 92 (1974), p. 3.

Bloom, Benjamin S. "Changing Conception of Examining at the University of Chicago." In *Evaluation in General Education*, edited by Paul L. Dressel. Dubuque, Iowa: William C. Brown Co., 1954.

Blum, Paul Von. "Needed: A Code of Ethics for Teachers." *Chronicle of Higher Education*, October 21, 1974, p. 20.

California State University. "Student Information." Unpublished. Los Angeles.

Carver, Ronald P. "Two Dimensions of Testing. Psychometric and Edumetric." *American Psychologist* 29 (1974): 512-518.

Collins, Janet E. and Nickel, K. N. "Grading, Recording and Averaging Practices in Higher Education." Mimeographed. Wichita, Kansas: Wichita State University, 1974.

Coulter, E. M. *College Life in the Old South.* 2d ed. Athens: The University of Georgia Press, 1951.

Cureton, Louise W. "The History of Grading Practices." *Measurement in Education*, no. 4 (1971), pp. 1-8.

Defunis v. Odegaard, 416 U.S. 312, 94 S. Ct. 1704 (1974).

Dressel, Paul L. *Evaluation in Higher Education.* Boston: Houghton Mifflin, 1961.

Ebel, Robert L. *Essentials in Educational Measurement.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.

———. "Writing the Test Item." In *Educational Measurement*, edited by E. F. Lindquist. Washington, D.C.: American Council on Education, 1966.

Etzioni, Amitai. "Grade Inflation: Neither Freedom nor Discipline." *Human Behavior*, October 1975, p. 11.

Felker, Daniel B., and Dapra, Richard A. "Effects of Question Type and Question Placement on Problem-Solving Ability from Prose Material." *Journal of Educational Psychology* 67 (1975): 380-384.

Fields, Cheryl M. *Chronicle of Higher Education*, April 29, 1973, p. 1.

Griggs v. Duke Power Company, 401 U.S. 424 (1971).

Group for Human Development in Higher Education. *Faculty Development in a Time of Retrenchment.* New Rochelle, N.Y.: *Change* Magazine, 1974.

Hechinger, Fred M. "An Academic Counter-Revolution." *Saturday Review/World*, no. 131 (1974), pp. 63-68.

Hofstadter, R. *Anti-Intellectualism in American Life.* New York: Vintage Books, 1966.

Huff, Sheila, "Credentialing by Tests or by Degrees: Title VII of the Civil Rights Act and Griggs v. Duke Power Company." *Harvard Educational Review*, no. 2 (1974).

Levine, Arthur and Weingart, John. *Reform of Undergraduate Education.* San Francisco: Jossey-Bass, 1973.

Lindquist, E. F. *A First Course in Statistics: Their Use and Interpretation in Education and Psychology.* Boston: Houghton Mifflin, 1942.

Lorge, Irving. "The Fundamental Nature of Measurement." In *Educational Measurement*, edited by E. F. Lindquist. Washington, D.C.: American Council on Education, 1966.

Mager, Robert F. *Goal Analysis*. Belmont, Ca.: Fearon, 1972.

———. *Measuring Instructional Intent*. Belmont, Ca.: Fearon, 1973.

McClelland, David C. "Testing for Competence Rather Than for 'Intelligence.'" *American Psychologist* 28 (1973): 1-14.

McGuire, Christine H. "An Evaluation Model for Professional Education-Medical Education." *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1968.

Meyer, G. "An Experimental Study of the Old and New Types of Examination." *Journal of Educational Psychology* 26 (1935): 30-40.

Michigan State University. "Code of Teaching Responsibility." Unpublished. East Lansing: 1969.

Pomona College. "Policy on Disputed Grades." Unpublished. Claremont, Ca.

Popham, W. James, ed. *Criterion-Referenced Measurement: An Introduction*. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Rhodes, Douglas W. *The Undergraduate Program for Counseling and Evaluation*. Princeton, N.J.: Educational Testing Service, 1973.

Ruskin, Robert S. and Hess, John H. *The Personalized System of Instruction in Higher Education: An Annotated Review of the Literature*. Washington, D.C.: Center for Personalized Instruction, Georgetown University, 1974.

Southern Regional Education Board. "Learning Your CBC's." *Regional Spotlight*. Atlanta: Southern Regional Education Board, September 1974.

Stalnaker, John M. "The Essay Type of Examination." In *Educational Measurement*, edited by E. F. Lindquist. Washington, D.C.: American Council on Education, 1966.

State Ex Rel. Bartlett v. Pantzer, 489 P. 2d 375 (1971).

Stice, James. "Progress Report on the PSI Project at the University of Texas at Austin." *PSI Newsletter*, no. 3 (1975).

Swarthmore College. "Critique of a College." Swarthmore, Pa.: 1967.

Swarthmore College Faculty. *An Adventure in Education*. New York: Macmillan, 1941.

Thomas, L. and Augstein, S. *An Experimental Approach to Learning From Written Material*. Oxbridge, England: Centre for the Study of Human Learning. Brunel University, 1970.

Warren, J. R. *College Grading Practices: An Overview*. Washington, D.C.: ERIC Clearinghouse on Higher Education, 1971.

Western Michigan University. "Student Academic Rights: Policies and Procedures." Unpublished. Kalamazoo.

Yeasmeen, Nazma and Barker, Donald G. "A Half Century of Research on Essay Testing." *Improving College and University Teaching*, no. 1 (1973).

Suggested Readings

American Psychological Association. *Standards for Educational and Psychological Tests.* Washington, D.C.: American Psychological Association, 1974.

This monograph was developed by a joint committee of members from the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. The contents are directed to both developers and users of standardized tests. "Essential," "very desirable," and "desirable" considerations about tests are proposed.

Anderson, Scarvia, Ball, Samuel, Murphy, Richard T., and Associates. *Encyclopedia of Educational Evaluation.* San Francisco: Jossey-Bass, 1975.

This is one of the first detailed reference works on concepts and techniques for evaluating education and training programs. It is not limited in scope to colleges and universities. The articles—alphabetically arranged from "accountability" to "variance"—are written by specialists. Each article is extensively cross-referenced and is followed by selected sources. The articles cover 11 topics: evaluation models; functions and targets of evaluation; program objectives and standards; social context of evaluation; planning and design; systems technologies; variables; measurement approaches and types; technical measurement considerations; reactive concern; analysis and interpretation.

Bowen, Howard R., ed. *New Directions for Institutional Research: Evaluating Institutions for Accountability,* No. 1. San Francisco: Jossey-Bass, Spring 1974.

As the title implies, this booklet is about program evaluation. The seven papers, prepared especially for this volume by six authorities, deal with the various complexities of assessment and offer suggestions for resolving them.

Bruning, J. L. and Kintz, B. L. *Computational Handbook of Statistics.* Glenview, Ill.: Scott, Foresman and Co., 1968.

This is an excellent "cookbook" of statistical methods, clear and concise in its presentation of the steps necessary to compute the basic measurement statistics mentioned in *The Testing and Grading of Students.*

Buros, Oscar K., ed. *The Seventh Mental Measurements Yearbook.* Highland Park, N.J.: Gryphon Press, 1972.

This work is in two volumes that have a total of slightly more than 2,000 pages. More than 1,100 published tests (achievement, attitude, personality, and others) are listed, along with some 12,000 references. For approximately half of the tests, there are orig-

inal reviews by experts, and there are around 200 reviews excerpted from journals. These volumes are indispensable when selecting a standardized test for either classroom use or research purposes.

Dressel, Paul L. and Associates. *Evaluation in Higher Education.* Boston: Houghton Mifflin, 1961.

This is one of the few books in this field beamed directly to college and university faculty members. Thus the level of discourse is more appropriate than that in many other tomes and examples of test questions tend to be quite practical. Of the 13 chapters, all written by different authorities, 10 deal explicitly with the issues discussed in *The Testing and Grading of Students.* Four of them are especially pertinent: evaluation in the social sciences; evaluation in the natural sciences; evaluation in the humanities; and evaluation of communication skills.

Ebel, Robert L. *Essentials of Educational Measurement.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.

This book, a revised version of the author's 1963 *Measuring Educational Achievement*, is sound, readable, and practical. It is referred to repeatedly throughout the first three chapters of the present work, and many points only touched on here are clearly elaborated therein. The 22 chapters are separated into five categories: Part I—History and Philosophy; Part II—Classroom Test Development; Part III—Getting, Interpreting, and Using Test Scores; Part IV—Test Analysis and Evaluation; Part V—Published Tests and Testing Programs. There is a glossary of the terms and concepts used in educational measurement.

Lindquist, E. F., ed. *Educational Measurement.* Washington, D.C.: American Council on Education, 1951.

This useful book, which went into its sixth printing in 1966, is a comprehensive handbook and textbook on the theory and technique of educational measurement. All 18 articles were especially prepared for the volume by noted authorities. Many of the selections—which are grouped into three categories, The Functions of Measurement in Education, The Construction of Achievement Tests, and Measurement Theory—are of a very practical nature, and all instructors can find good tips here for testing.

Mager, Robert F. *Goal Analysis.* Belmont, Ca.: Fearon, 1972.

Mager's work merits considerable attention. His writing is clear and easily understood; he comfortably translates his theory into application. *Goal Analysis* is a small book (136 pages) that spells out the steps by which instructors can identify goals in their instruction and establish the appropriate steps toward the successful completion of those goals. Assessment and evaluation are both built into the goal-analysis procedure. The book defines procedures that allow instructors to say where they are, where they want to go, how they intend to get there, and how they know when they are there.

Mager, Robert F. *Measuring Instructional Intent.* Belmont, Ca.: Fearon, 1973.

Writing in his unique, informal style, the author describes and illustrates a procedure that will help in selecting or creating test items that will match objectives. Illustrations cover a wide array of performances.

Mager, Robert F. *Preparing Instructional Objectives.* 2d ed. Belmont, Ca.: Fearon, 1975.

While the contents of this book seem deceptively simple, the substance is profound, especially for those who have given almost no thought to objectives. The book is cleverly and wittily written. Beginners in the academic enterprise will benefit greatly; old-timers might.

Mehrens, William A. and Ebel, Robert L., eds. *Principles of Educational and Psychological Measurement: A Book of Selected Readings.* Chicago: Rand McNally, 1967.

This book contains classical articles on measurements, most of them very technical and statistical, which were published over a span of 30 years. The 37 selections are grouped into five categories: measurement theory and scaling; norms; reliability; validity; item analysis and selection.

Pace, C. Robert, ed. *New Directions for Higher Education: Evaluating Learning and Teaching,* No. 4. San Francisco: Jossey-Bass, 1973.

Each chapter was prepared especially for this booklet by authors with widely varying perspectives. The six papers collectively demonstrate how complex problems of eval-

uation are and the innumerable factors to be considered and are useful as a quick but substantive overview.

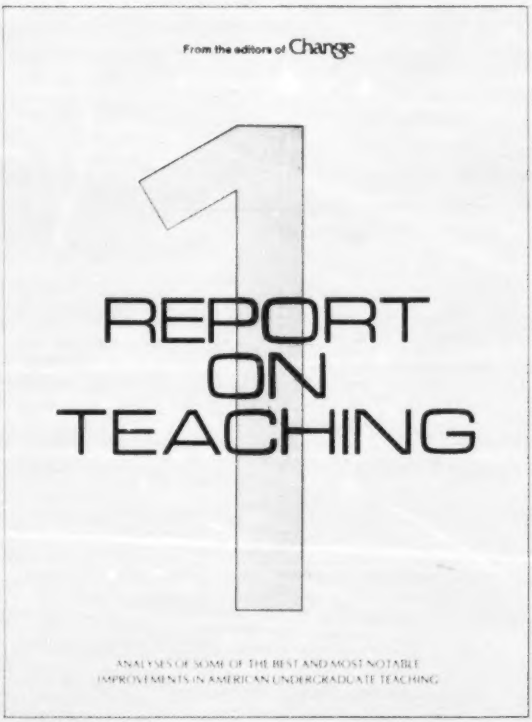
Thorndike, Robert L., ed. *Educational Measurement*. 2d ed. Washington, D.C.: American Council on Education, 1971.

The first edition of this book went through seven printings. This second edition, prepared with the assistance of the American Educational Research Association and the American Council on Education, reflects the broadened concern about evaluation that has been developing. The 20 pieces are addressed to four areas: Part One--Test Design, Construction, Administration, and Processing; Part Two—Special Types of Tests; Part Three—Measurement Theory; Part Four—Application of Tests to Educational Problems. Both the specialist and the novice will find this book useful.

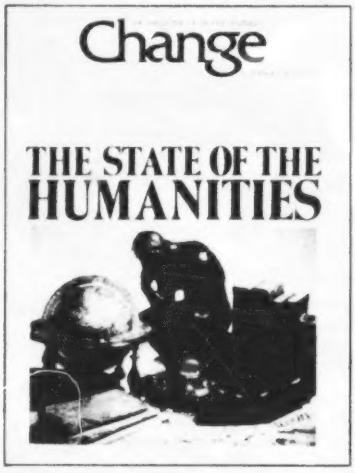
Selected Journals with Special Emphasis on Evaluation

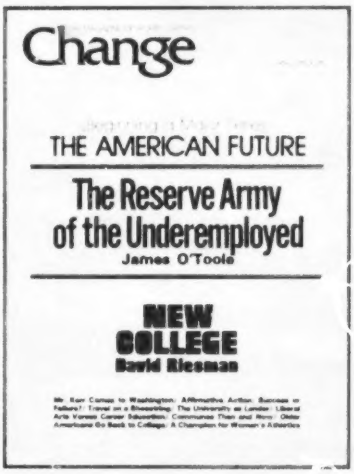
American Educational Research Journal
British Journal of Statistical and Mathematical Psychology
Center for the Study of Evaluation
College Student Journal
Educational and Psychological Measurement
Journal of Educational Measurement
Journal of Research in Science Teaching
Programmed Learning and Educational Technology
Psychometrika
Review of Educational Research

On The Quality Of Teaching



In collaboration with the major disciplinary fields, *Change* is now publishing selected assessments of exceptional teaching on a twice yearly basis. Three fields of study are surveyed in each semi-annual issue. These Reports are made available through a grant from the Fund for the Improvement of Postsecondary Education. Please send your requests on official letterhead, along with $1 per copy to cover postage and handling, to Undergraduate Teaching Program, *Change*, NBW Tower, New Rochelle, N.Y. 10801.





The One Magazine for Academic People

Change is the first and only magazine to forge exciting new bonds among academics everywhere, regardless of their field and interest. Each month, *Change*'s 80,000 readers share in some of the most challenging editorial fare available. *Change* is the venturesome magazine of creative ideas, of major essays written by some of America's great minds, and ten regular features each month that are worth the price of subscription alone.

Change not only interprets a changing culture; it helps create it. For those who thrive on more than yesterday's news, reading *Change* can be a revealing experience. Use the handy order form in the front of the book or send for a one-year subscription for $14 to: *Change* Magazine, NBW Tower, New Rochelle, N.Y. 10801.

Some 1975 editorial highlights

AMERICANA

Christopher Lasch on The Democratization of Culture ☐ Orlando Patterson on Ethnicity ☐ Victor Navasky on the Encyclopaedia Britannica ☐ Edwin Newman on How Academics Kill the English Language ☐ James Degnan on Jessica Mitford ☐ James Real on the Center for the Study of Democratic Institutions.

SOCIAL ISSUES

James O'Toole on The Reserve Army of the Underemployed ☐ Kenneth Boulding on the Management of Decline ☐ Richard Lester on the Equal Pay Boondoggle ☐ Cynthia Secor on Lesbians in Academe ☐ John Egerton on Adams v. Richardson ☐ Seymour Martin Lipset on Harvard's Economics Department ☐ Marilyn Gittell on the Failure of Affirmative Action ☐ Daniel Greenberg on The Politics of Science.

THE WORLD OF EDUCATION

David Riesman on New College ☐ Richard Freeman and Herbert Hollomon on the Declining Value of College Going ☐ Angela Stent on the Radcliffe Institute ☐ Peter M. Blau and Rebecca Margulies on America's Leading Professional Schools ☐ Who's Who in Higher Education ☐ Robert Lekachman on the Academic Labor Market ☐ Barry Mitzman on Union Power in Academe ☐ Arnold Sawislak on John Brademas.

ARTS & LETTERS

The Future of the Humanities ☐ Cullen Murphy on Campus Best-Sellers ☐ Gerald Holton on the Humanistic Basis of Scientific Work ☐ Joan Baum on Anthony Burgess's Clockwork Testament ☐ Sara Blackburn on The Academic Novel ☐ Harold Taylor on Student Expression ☐ Vermont Royster on The New Illiteracy ☐ Easy Klein on the Custom-Made Textbook.